인사이트

웨이스 시험 설비에서 AI 모델은 설비가 멈춘 원인을 단순한 규칙보다 적게 맞혔습니다

방동걸 (주)웨이스 대표이사
2026년 10월 12일

설비가 멈추면 PLC는 고장 코드 하나를 화면에 띄웁니다. 그런데 같은 코드 뒤에 원인이 여럿일 수 있습니다. 작업자는 그 자리에서 무엇부터 확인할지 정해야 하고, 요즘은 이 일을 AI에게 맡기자는 이야기가 많습니다.

제조 AI 운영체제 VEXPLOR를 만드는 웨이스는 이 생각이 맞는지 시험 설비에서 확인했습니다. PLC 고장 코드만으로 원인을 하나로 정할 수 없는 165문항에서, 추가 학습 없는 4B 언어 모델은 40문항을 맞혔고 고장 코드로 후보를 좁힌 단순한 규칙은 90문항을 맞혔습니다. 웨이스는 이 결과를 2026년 10월 8일 기술 보고서 TR-2026-08로 공개했습니다. 이 글에는 실험 방법과 알게 된 것, 아직 모르는 것을 차례로 적습니다.

01

웨이스는 시험 설비를 2,997번 실행하며 고장을 일부러 넣었습니다

실제 공장에서는 설비가 멈춘 원인의 정답을 알기 어렵습니다. 작업자가 고친 기록은 남아도, 정말 그것이 원인이었는지는 다시 확인하기 어렵습니다. 그래서 웨이스는 정답을 아는 가상 시험 설비를 만들었습니다.

시험 설비는 셀 3개로 되어 있습니다. 그중 셀 B에는 벨트 컨베이어, 검사 장치, 도착 센서, 상자를 슈트로 밀어내는 공압 푸셔가 있습니다. 생성기가 고장을 하나, 둘, 셋씩 섞어 넣고 서서히 닳는 마모도 넣었습니다. 실행은 2,997건, AI에게 물을 판단 구간은 20,016개였습니다. 멈춘 원인의 정답은 넣은 고장을 하나씩 빼고 다시 실행해 보고 정했습니다.



웨이스 시험 설비는 고장을 일부러 넣어 멈춘 원인의 정답을 알고 시작합니다

과제	AI에게 준 기록	정답을 정한 방법
멈춘 원인	멈추기 직전 10초	고장을 하나씩 빼고 재실행
복귀 절차	멈춘 뒤 10초	복귀 절차표
정비 징후	마모 추세 최대 300초	넣은 마모 수준
알람 우선순위	2초 안의 알람 묶음	알람 우선순위 표

AI 모델은 기록을 읽고 추천만 했습니다. PLC에 쓰는 연결은 없었고, 래더·안전 릴레이·비상정지 판단은 그대로 두었습니다. 판정 기준은 실험 전인 10월 6일에 문서로 고정했고, 결과를 본 뒤 바꾼 기준은 없습니다. 다만 고정 기록은 사내 저장소에만 남겼고 공개는 결과를 본 뒤 정했기 때문에, 외부에서 이를 제3자 기록으로 확인할 수는 없습니다.

02

AI 모델은 규칙이 못 정한 문항에서 규칙보다 30%포인트 낮았습니다

비교 대상인 규칙은 단순합니다. PLC가 처음 띄운 고장 코드를 보고 가능한 원인 후보를 고릅니다. 후보가 하나면 그것을 답하고, 여럿이면 학습용 기록에서 가장 많았던 원인을 답합니다.

AI 모델은 Qwen 3.5 4B를 추가 학습 없이 썼고, 선택지 가운데 한 글자로 답하게 했습니다. 모델에게는 고장 코드를 가린 기록을 주었습니다. 코드를 보여 주면 기록을 읽지 않고도 답할 수 있으니, 모델이 기록에서

원인을 찾아내는지 보려고 가린 것입니다. 규칙이 원인을 하나로 정한 문항은 규칙의 답을 쓰고, 규칙이 못 정한 문항만 모델의 답으로 바꿨습니다.

두 쪽이 본 기록은 같지 않았습니다. 규칙은 고장 코드로 후보를 2~3개로 좁혔고, 그 후보에는 165문항 가운데 160문항에서 정답이 들어 있었습니다. 모델은 코드 없이 셀 B의 원인 범주 8개 전부에서 골랐습니다. 그래서 보고서는 사전 등록에 없던 추가 분석으로, 모델에게도 규칙의 후보 안에서만 고르게 했을 때를 따로 계산했습니다. 그때도 모델은 85문항을 맞혀 규칙의 90문항에 못 미쳤습니다.



고장 코드만으로 원인을 정할 수 없는 165문항에서 규칙은 90문항, AI 모델은 40문항을 맞혔습니다

셀 B 문항	규칙만	규칙 + AI 모델
규칙이 정한 741문항	700 (94.5%)	700 (규칙 그대로)
규칙이 못 정한 165문항	90 (54.5%)	40 (24.2%)
전체 906문항	790 (87.2%)	740 (81.7%)

규칙의 54.5%도 높은 값은 아닙니다. 후보 가운데 자주 나왔던 쪽을 고른 것뿐입니다. AI 모델은 그 단순한 기준에도 30%포인트 못 미쳤고, 모델을 붙이자 셀 B 전체 정답률은 87.2%에서 81.7%로 오히려 내려갔습니다.

03

AI 모델은 선택지 순서만 바꿔도 다른 답을 냈습니다

웨이스는 문항마다 선택지 순서를 바꿔 최대 5번 물었습니다. 규칙이 못 정한 165문항 가운데 148문항(90%)에서, 순서를 바꾸자 모델이 가장 높게 고른 답이 달라졌습니다. 모델이 답한 전체 10,387문항으로 넓혀도 8,404문항(81%)이 그랬습니다. 틀린 답을 고르면서 선택 확률을 0.9 이상으로 낸 문항도 93개였습니다.

셀 B에서 비상정지나 라이트 커튼 같은 안전 기능이 제대로 작동해 멈춘 192건 가운데 규칙은 191건을 맞혔고, 고장 코드를 가린 기록만 읽은 모델은 89건을 맞혔습니다. 규칙은 안전 정지를 가리키는 고장 코드를 그대로 읽으므로 같은 조건의 비교는 아닙니다. 다만 모델은 기록만으로는 안전 정지 192건 가운데 103건을 다른 원인으로 답했습니다.

반대로 규칙이 하나도 맞히지 못한 기계 걸림 54건에서는 모델이 46건을 맞혔습니다. 다만 모델은 기계 걸림이 아닌 문항 347개에도 기계 걸림이라고 답했습니다. 원인을 알아봤다고 볼 근거는 없고, 그 답을 자주 고른 결과로 보입니다.

선택지를 늘어놓는 순서처럼 원인과 상관없는 조건에 따라 모델의 답이 바뀐다는 뜻입니다. 보고서는 이런 문항을 사람이 다시 확인할 문항으로 분류했습니다.

04

정비 징후는 정답을 만든 경계값을 그대로 쓴 규칙이 53건 중 51건을 잡았습니다

설비가 멈추기 전 마모 징후를 읽는 과제에서는 최대 300초 동안의 행정 시간, 벨트 속도, 센서 신호 떨림(채터링) 추세를 주고 정상·관찰·정비 필요 가운데 고르게 했습니다. 규칙은 행정 시간 1.15배, 벨트 속도 편차 5% 같은 경계값으로 판정합니다.

정비 징후	규칙	AI 모델
정비 필요 실행 53건	51건 잡음	53건(전부 정비 필요)
정비 필요 아닌 108건 오탐	0건	108건 전부
수준 정답률(구간 412개)	94.9%	39.1%

셀 B에는 벨트 속도 경고가 없어서, PLC가 끝까지 경고를 내지 않는 벨트 이상 실행 19건을 일부러 넣었습니다. 규칙은 19건을 모두 잡았고, 모든 실행을 정비 필요로 낸 모델도 19건을 모두 잡았습니다. 이

결과를 규칙이 이겼다는 근거로 읽으면 안 됩니다. 이 규칙은 정답을 만든 경계값을 그대로 쓴 규칙이라 잘 맞는 것이 당연합니다. 실제 공장에서는 그 경계값을 어떤 기록으로 정하느냐부터 따져야 합니다.

복귀 절차와 알람 우선순위도 사정이 같습니다. 정답이 곧 절차표와 우선순위 표라서 규칙이 맞는 것이 당연합니다. 같은 문항에서 모델의 정답률은 복귀 절차 14.4%, 알람 우선순위 54.2%였습니다.

05

웨이스의 이번 실험은 실제 공장 기록과 미세조정 모델을 시험하지 않았습니다

이 결과가 말하는 범위는 좁습니다. 웨이스는 보고서에 다음 한계를 함께 적었습니다.

- **시뮬레이션만 했습니다.** 시험 설비의 물리값은 가정이고, 실물 설비와 대조한 결과는 없습니다. 실제 공장 기록에서 어떻게 될지는 이 실험이 말하지 않습니다.
- **모델은 하나, 추가 학습은 없었습니다.** 공장 기록으로 미세조정한 모델은 시험하지 않았습니다. 결론은 「추가 학습 없는 4B 일반 모델이 한 글자로 답하는 방식」에 한합니다.
- **정비 징후의 모델 기준값을 고를 기록이 부족했습니다.** 검증용 기록에 정비 필요 실행이 2건뿐이라 모델의 기준값이 낮게 잡혔고, 모델은 모든 실행을 정비 필요로 판정했습니다. 다음 실험에서는 검증용 기록도 늘립니다.
- **규칙과 모델이 본 기록이 달랐습니다.** 규칙은 첫 고장 코드를 보고 모델은 보지 못했습니다. 모델을 규칙의 후보 안에 묶은 계산(85문항)은 사전 등록에 없던 추가 분석입니다.
- **사전 등록은 사내 저장소에만 있습니다.** 판정 기준은 실행 전에 고정했지만, 그 기록은 비공개이고 공개 여부는 결과를 본 뒤 정했습니다.
- **사람 심사를 받지 않았습니다.** 도구·채점기·문안은 AI 코딩 에이전트가 만들었고 검수도 같은 계열 에이전트가 했습니다.

그래서 이 글의 결론은 「AI는 공장에 쓸모없다」가 아닙니다. 이 조건에서 이 모델은 단순한 규칙을 넘지 못했고, 그것을 확인하려면 규칙 기준선이 먼저 있어야 했다는 것입니다.

06

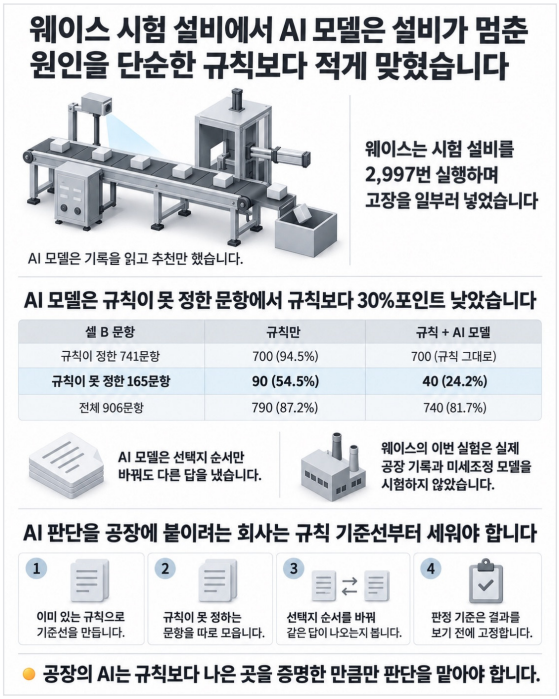
AI 판단을 공장에 붙이려는 회사는 규칙 기준선부터 세워야 합니다

고장 코드를 읽는 규칙 기준선이 없었다면, 모델의 24.2%가 현장에서 이미 쓰는 방식보다 나은지 알 수 없었습니다. 실험에서 얻은 순서를 4가지로 정리합니다.



AI 판단을 붙이기 전에는 규칙 기준선, 규칙이 못 정한 문항, 순서 바꿔 묻기, 기준 고정을 차례로 갖춥니다

1. **이미 있는 규칙으로 기준선을 만듭니다.** 고장 코드표, 알람 우선순위 표, 마모 경계값이 그대로 기준선이 됩니다. AI는 이 기준선보다 나을 때만 붙입니다.
2. **규칙이 못 정하는 문항을 따로 모읍니다.** AI가 값을 보태야 할 곳은 규칙이 이미 맞히는 곳이 아니라 거기입니다. 이번 실험에서는 165문항이었습니다.
3. **선택지 순서를 바꿔 같은 답이 나오는지 봅니다.** 순서에 따라 답이 바뀌는 추천은 작업자 화면에 띄우지 않습니다.
4. **판정 기준은 결과를 보기 전에 고정합니다.** 결과를 본 뒤에 기준을 고르면 어떤 모델이든 좋아 보일 수 있습니다.



이 글을 한 장으로 요약하면 이렇습니다

웨이스는 VEXPLOR에서 AI의 판단을 PLC로 보내기 전에 가상 공장에서 먼저 실행하도록 개발하고 있습니다. 이번 실험은 그 앞 단계, 곧 AI가 내는 추천이 기존 규칙보다 나은지를 먼저 확인하는 일이었습니다. 보고서 원문과 결과 파일·검산 스크립트 묶음은 wace.me 기술 보고서와 Zenodo(DOI 10.5281/zenodo.23225139)에서 받을 수 있습니다.

공장의 AI는 규칙보다 나은 곳을 증명한 만큼만 판단을 맡아야 합니다. WACE(웨이스)는 공장이 스스로 판단하고 멈추지 않게 하는 제조 AI 운영체제 VEXPLOR를 만드는 회사다.

07

함께 읽어보세요

- 웨이스가 VEXPLOR에서 AI의 판단을 PLC로 보내기 전에 같은 스캔의 가상 공장에서 먼저 실행하려는 이유
- 웨이스가 부정적 결과로 끝난 AI 명령 검증 실험을 기술보고서 TR-2026-01로 공개했습니다
- 설비 제어 로직을 소프트웨어로 실행해 보면, 처음 틀리는 곳은 3군데입니다

원문 blog.vexplor.com/post/ai-model-vs-rule-plc-stop-cause
발행 2026년 10월 12일 · 방동걸 (주)웨이스 대표이사

회사 주식회사 웨이스 (WACE Inc.) · 대표이사 방동걸 · 사업자등록번호 439-87-02852
주소 (18462) 경기 화성시 동탄구 동탄영천로 150, 에이동(A동) 1702-1708호 (영천동, 실리콘앨리동탄)
연락처 전화 070-4730-0629 · contact@wace.me
홈페이지 wace.me · linkedin.com/company/wace-inc

원문 홈페이지 링크드인